Two Generalizations of the Design of Experiments Methodology for Enhanced Process Understanding

Christos Georgakis

Department of Chemical and Biological Engineering

and Systems Research Institute

TUFTS University, Medford, MA 02155, USA

Machine Learning Algorithms for Chemical Reaction Systems

Christos Georgakis

Department of Chemical and Biological Engineering

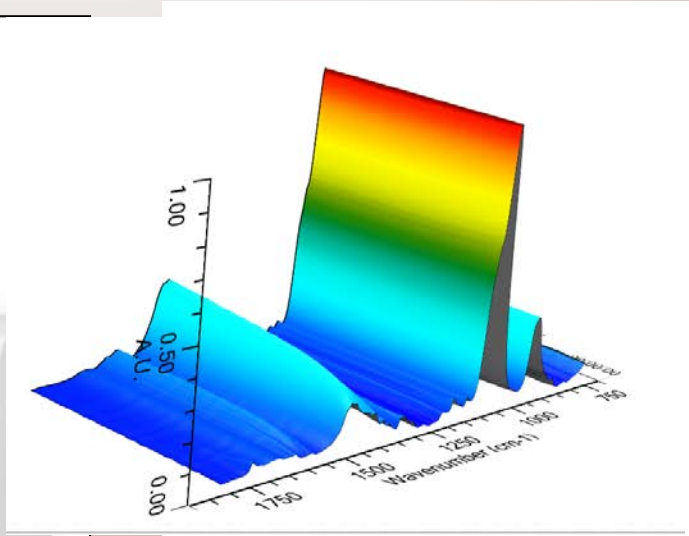
and Systems Research Institute

TUFTS University, Medford, MA 02155, USA

Types of Models

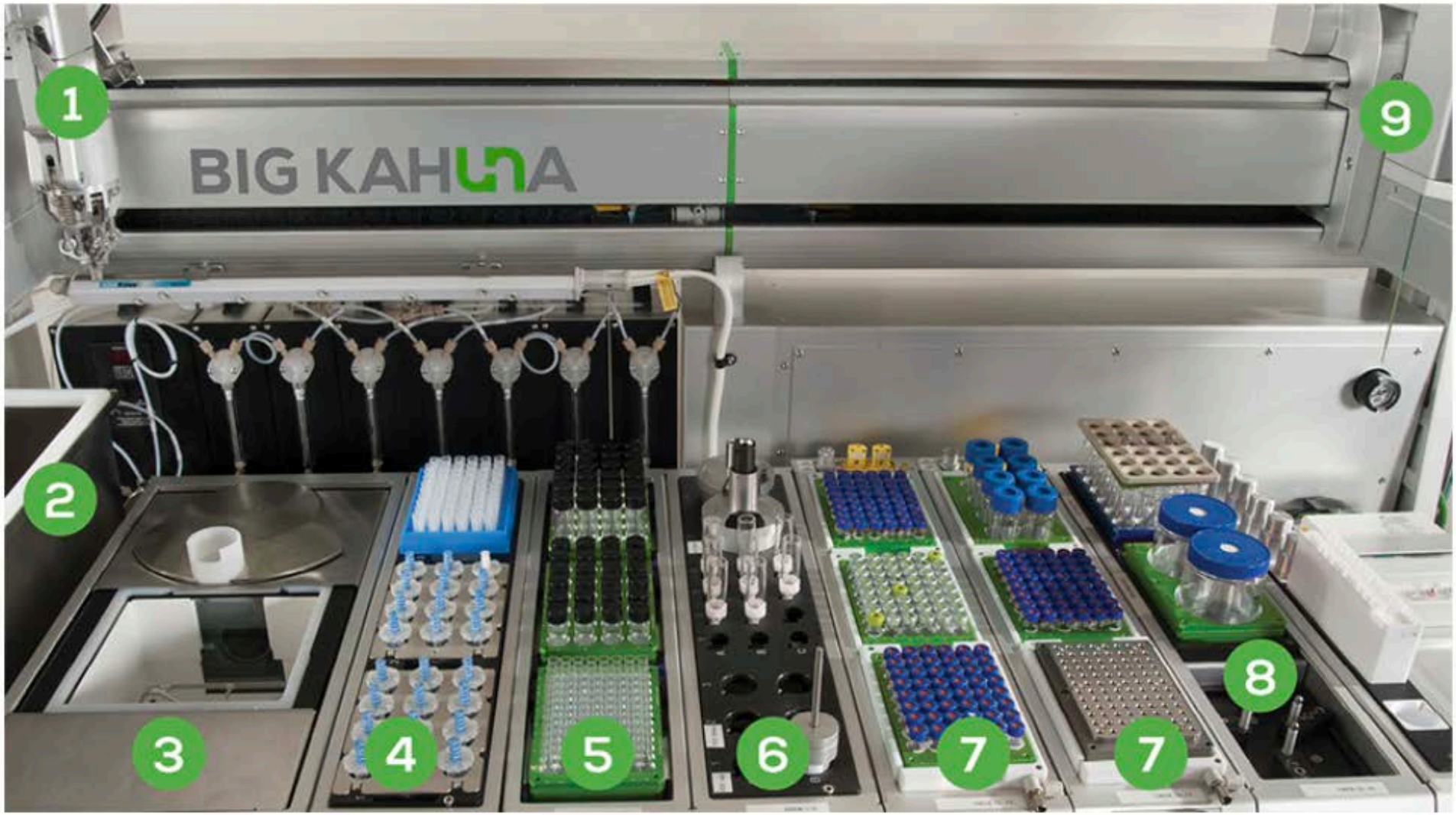
- Knowledge-Driven Models
  - Instead of "Fundamental" or "First Principles"
- Data-Driven Models
  - No knowledge of Inner Workings of Process
- Hybrid Models
  - Partial Knowledge + Data
- ***Models Should Have a Purpose***
  - Change the Purpose → Change the Model
    - Models for Kinetics, Process Design, Optimization, Control …
    - Conceptual, Physical (Pilot Plant), Mathematical, …

Plethora of Robotic Devices



If You have One … Million $

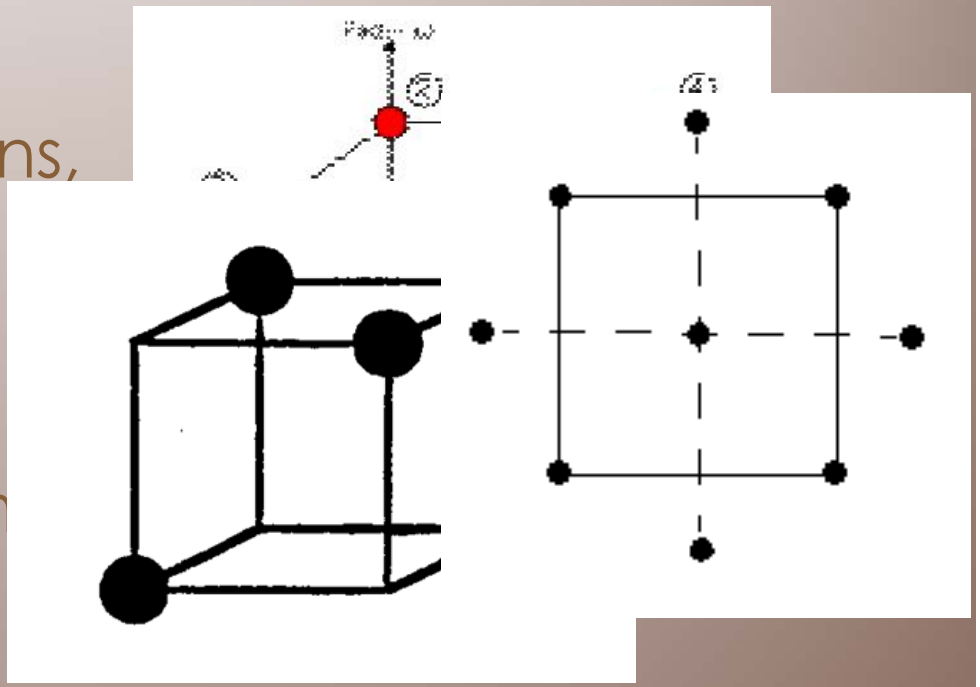
The Age of Big Data



Design Experiments - Analyze Data

Design of Experiments (DoE)

- Very Powerful Methodology 50 Years Young!
- ***Never Change One Condition at a time***
- Full Factorial Designs,
- Fractional Factorial Designs,
  - ½ fraction: 2^{n-1}
  - ¼ fraction: 2^{n-2}
  - 1/8 fraction: 2^{n-3}
- Center Composite Design
  - …



Is DoE Sufficient?

My Answer is: **NO**

1st Generalization: DoE ➔ DoDE

- Time-Varying Inputs (Factors)
  - Design of Dynamic Experiments (DoDE)
    - Batch Reactor Temperate vs. Time, T(t)=?
    - Feeding of Bioreactor with Sugar Source, u(t)=?
    - Bioreactor pH vs. Time, pH(t)=?
  - How Many Dynamic Experiments?
  - How we Design them?

Georgakis, C., (2013) *"Design of Dynamic Experiments: A Data-Driven Methodology for the Optimization of Time-Varying Processes"* Ind. Eng. Chem. Res. **52** (35):12369-12382

2nd Generalization: RSM ➔ DRSM

- RSM: Response Surface Methodology
  - Interpolative Polynomial Model of Output
  - $y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{ij} X_i X_j + \sum_{i=1}^{n} \beta_{ii} X_i^2$
- Composition Measurements every Hour
  - For 12 hrs ➔ 12 RSMs ??
- **DRSM:** Dynamic Response Surface Method

$$y(t) = \beta_0(t) + \sum_{i=1}^{n} \beta_i(t) X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{ij}(t) X_i X_j + \sum_{i=1}^{n} \beta_{ii}(t) X_i^2$$

- $\beta_q(t) = \gamma_{q,1} P_0(t) + \gamma_{q,2} P_1(t) + \cdots \gamma_{q,R} P_{R-1}(t)$
  - $P_i(t)$ the i[th] Shifted Legendre Polynomial
  - $P_0 = 1,\ P_1(t) = -1 + 2t,\ P_2(t) = 1 - 6t + 6t^2$

DoDE: Time-Varying Inputs

- Define Time-Varying Input Domain
- Define Time-Varying Coded Variable, z(τ)

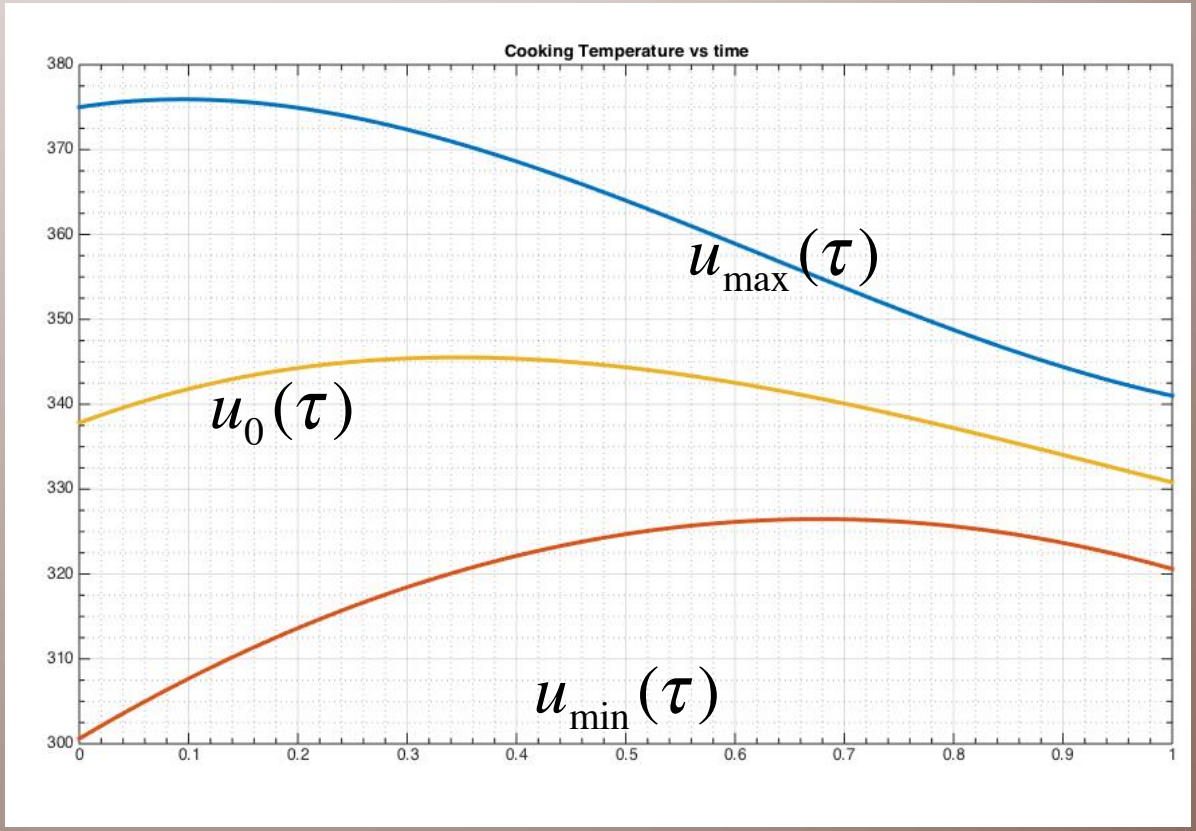
$$u_0(\tau) = \frac{u_{\max}(\tau) + u_{\min}(\tau)}{2}$$

$$\Delta u(\tau) = \frac{u_{\max}(\tau) - u_{\min}(\tau)}{2}$$

$$z(\tau) = \frac{u(\tau) - u_0(\tau)}{\Delta u(\tau)}$$

$$-1 \le z(\tau) \le +1, \quad \tau = t / t_b$$

$$u(\tau) \triangleq u_0(\tau) + \Delta u(\tau) z(\tau)$$

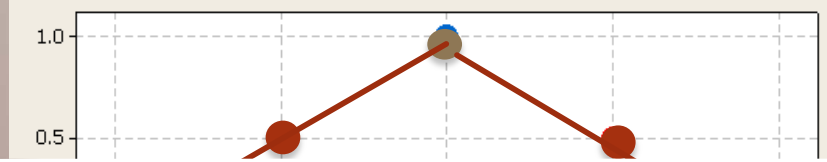


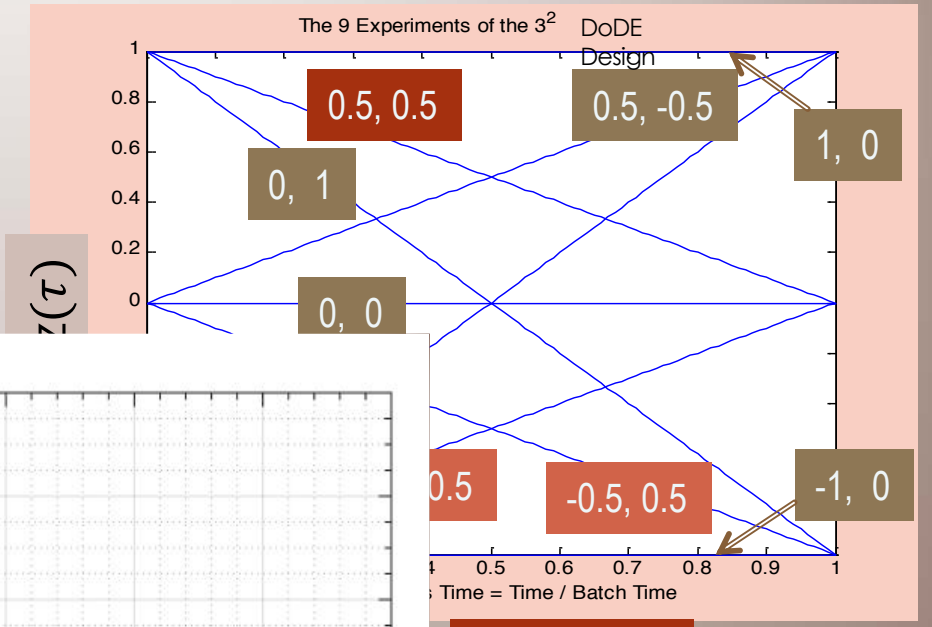
Main Idea: $z(\tau) = a_1 P_0(\tau) + a_2 P_1(\tau) + a_3 P_2(\tau) + \cdots$

Nine (9) Time-Varying Inputs

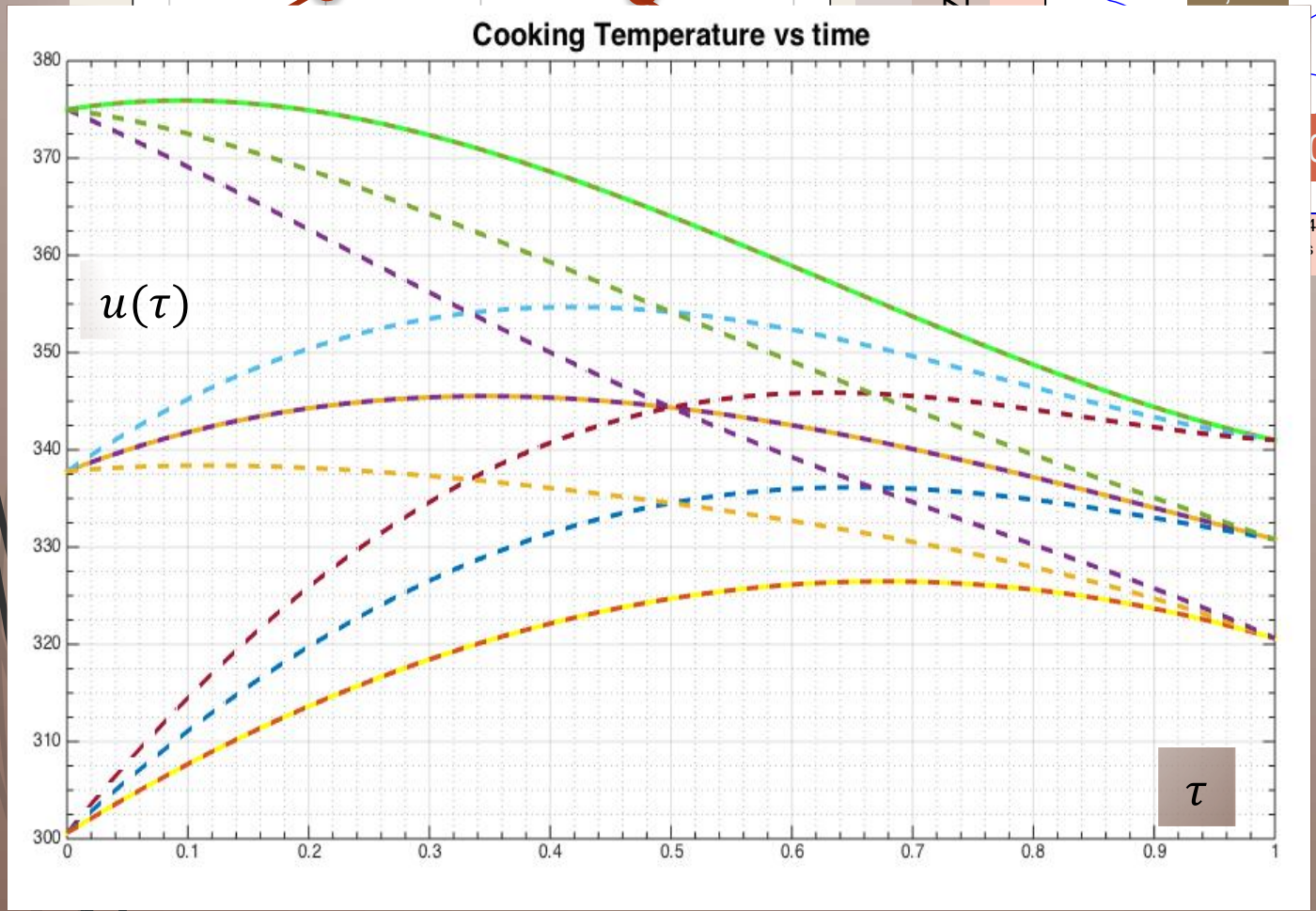
$$z(\tau) = a_1 P_0(\tau) + a_2 P_1(\tau)$$

$$-1 \le a_1 \pm a_2 \le +1 \Leftrightarrow -1 \le z(\tau) \le +1$$





$$\tau = t/t_b$$



DoDE Example: Batch Reactor

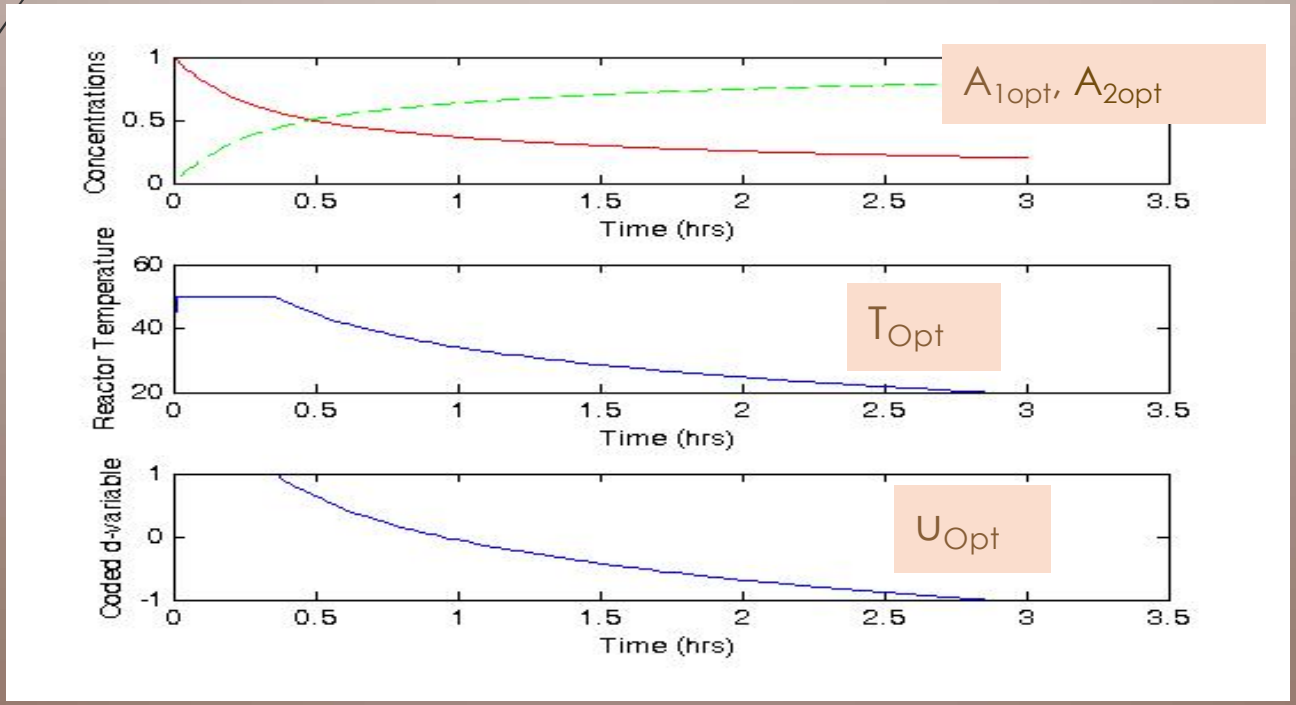
Reversible Reaction in Batch

$A_1 \rightleftharpoons A_2$ (15 < T < 50 °C)

$$\boldsymbol{r = k_1 A_1 - k_2 A_2} \qquad k_i = k_{i0} \exp\left(-\frac{E_i}{RT}\right) \; with \quad E_2 > E_1$$

Model-based Optimum Conversion: **74.6%**

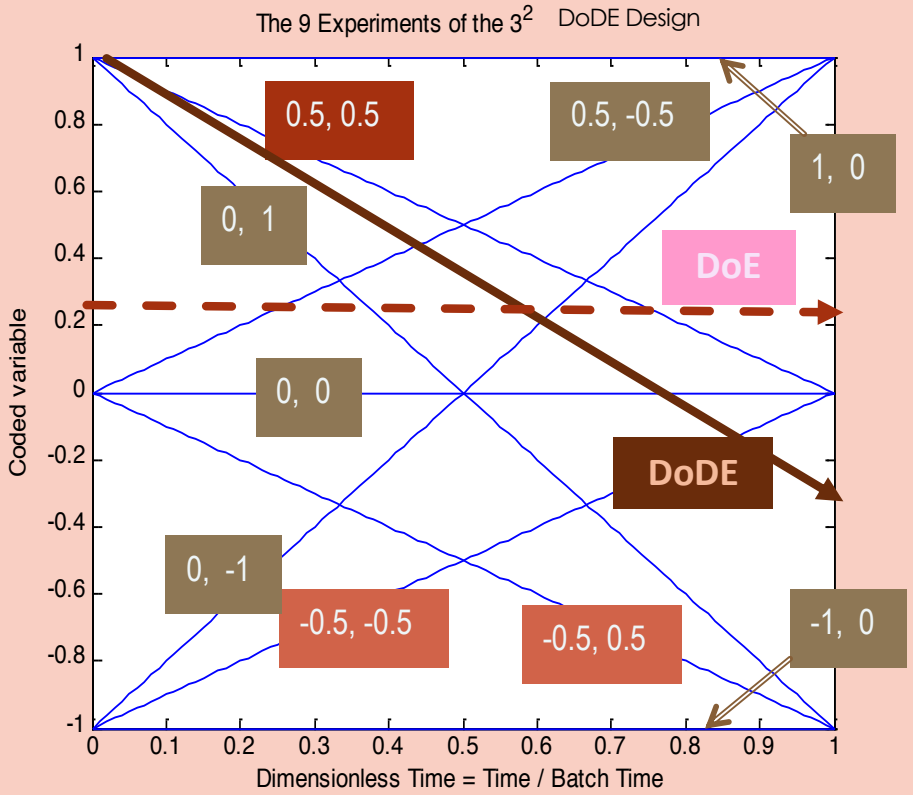
Decreasing Temperature Profile



Optimization via DoDE

- Two Factors: T Level & Linear Slope
  - Nine DoDE Experiments
    - Linear in Time
    - between 15^{0}C to 50^{0}C
  - Optimization:
    - Max DoDE Conversion
      - x=74.3%,
      - T* 50^{0}C → 28^{0}C

74.6%



	DoE	DoDE	MBO
Conversion	71.4%	74.3%	74.6%
Difference from MBO	3.2	0.3	

VERY Small Difference

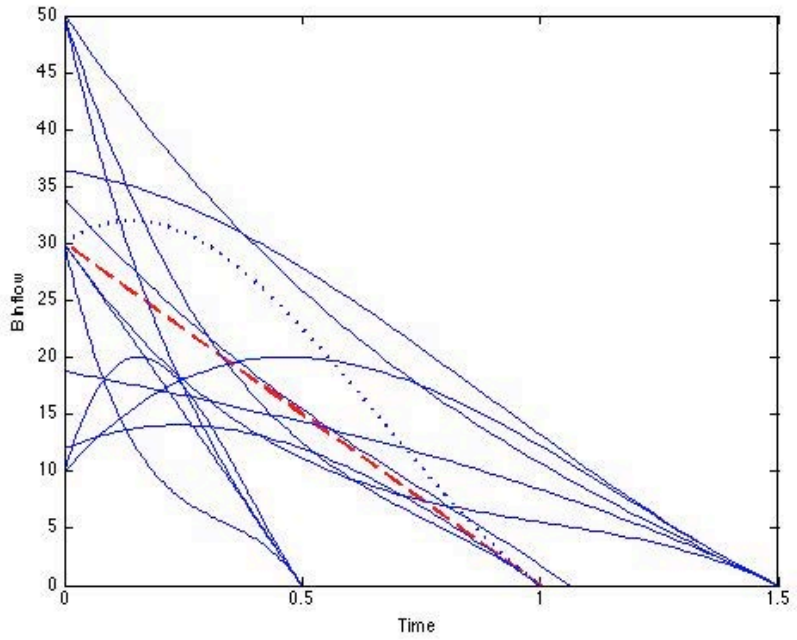
DoDE Semi-Batch Reactor

Tufts UNIVERSITY

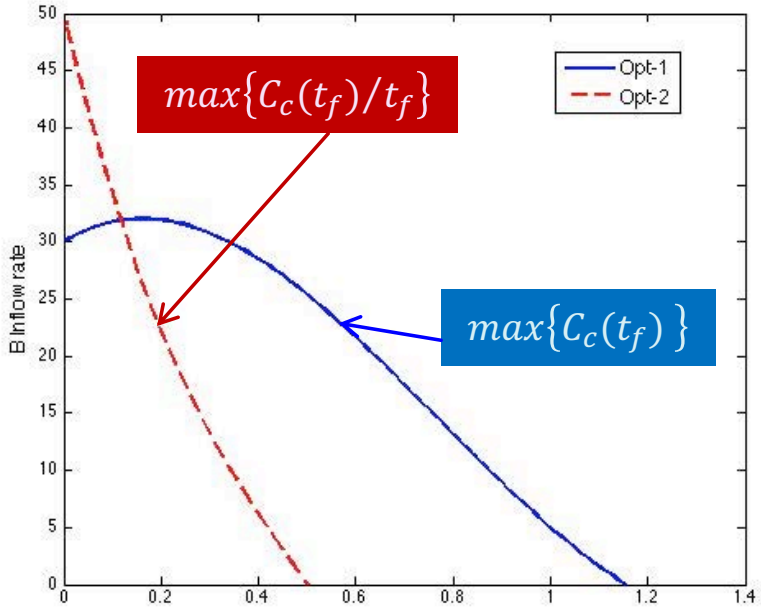
Reaction Example:

$$Rxn1:\ A + B \longrightarrow C \quad r_1 = k_1 C_A C_B, k_1 = 2\ l\ mol^{-1} h^{-1}$$
$$Rxn2:\ 2B \longrightarrow D, \quad r_2 = k_2 C_B^2,\ k_2 = 1\ l\ mol^{-1} h^{-1}$$
$$Rxn3:\ C \longrightarrow E, \quad r_1 = k_3 C_C,\ k_3 = 1\ h^{-1}$$

DoDE Runs: Feeding B

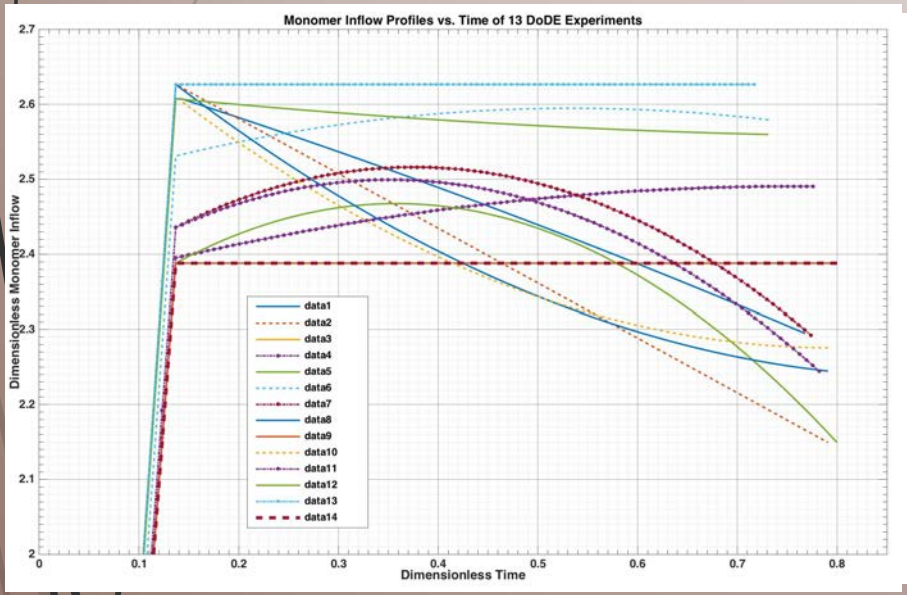


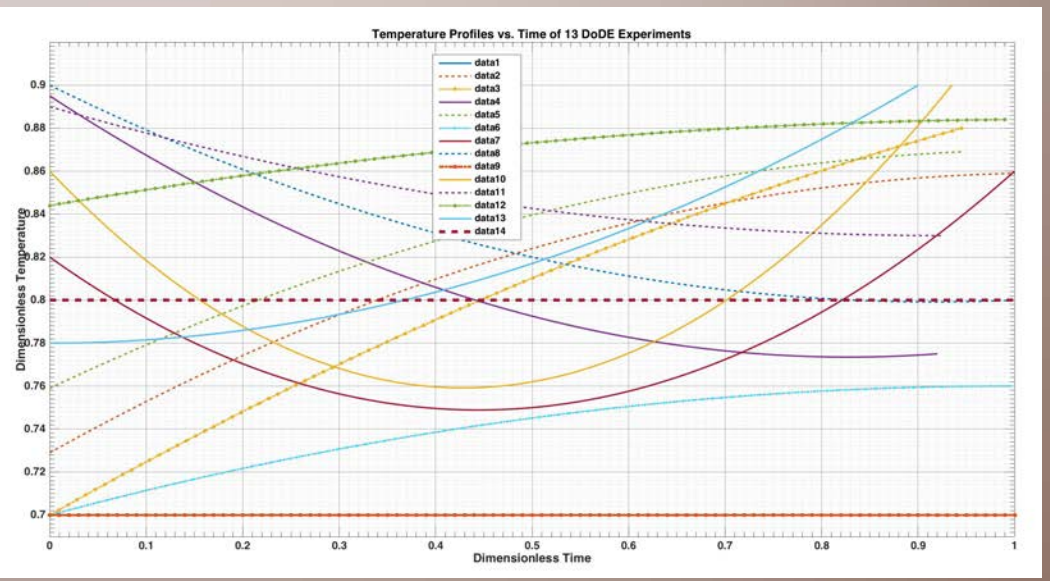
Optimal Runs



DoDE: The Dow Project

- Polymerization ➔ Increase Productivity
  - **NO** Detailed Knowledge-Driven Process Model
- Inputs (factors) Can Vary with Time





15 DoDE experiments ➔ Batch Time Reduced by 20%
Productivity Increase by 20%

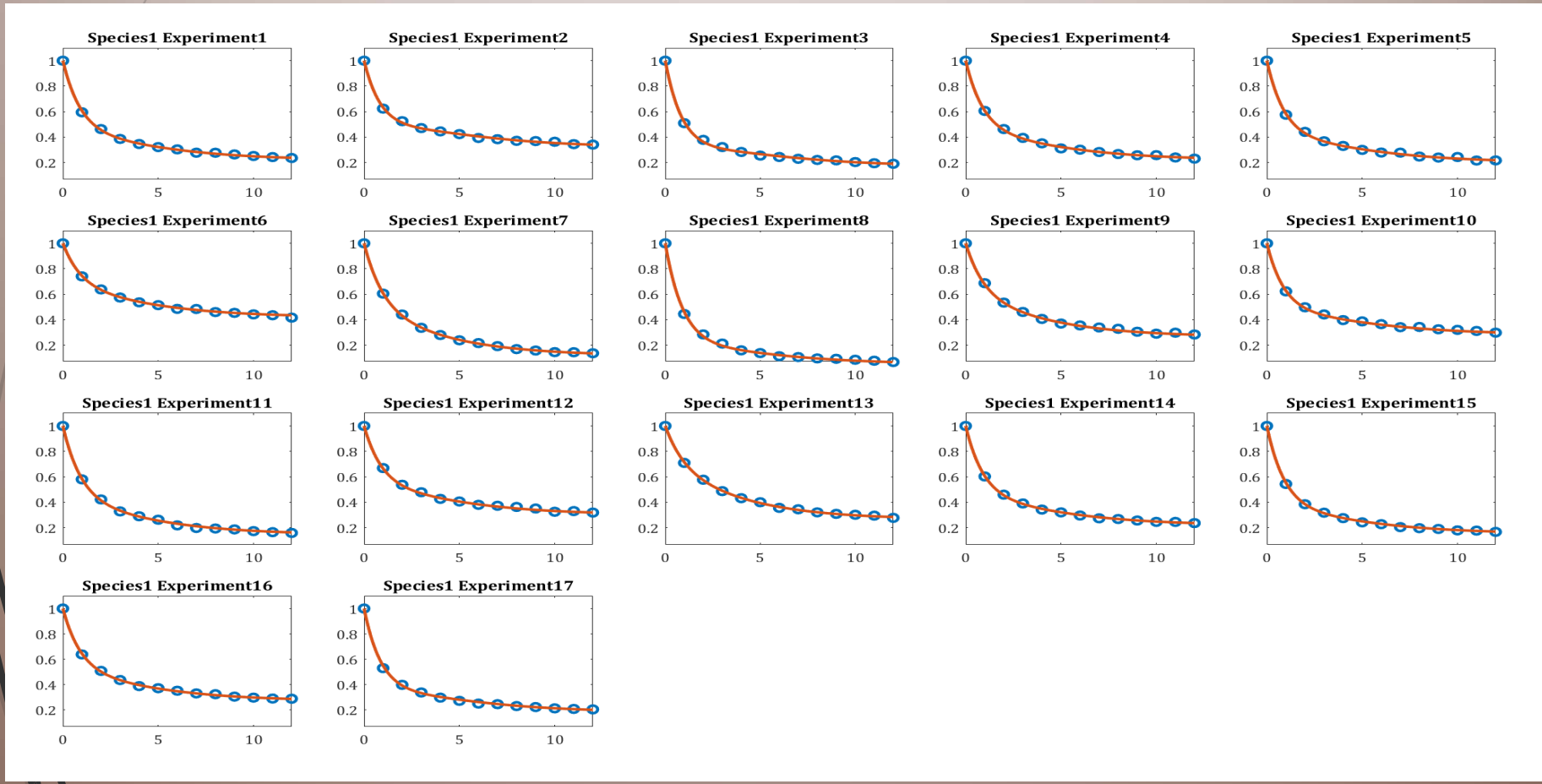
The DRSM Idea

- From RSM:
  - $y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{ij} X_i X_j + \sum_{i=1}^{n} \beta_{ii} X_i^2$
- To DRSM:
  - $y(t) = \beta_0(t) + \sum_{i=1}^{n} \beta_i(t) X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{ij}(t) X_i X_j + \sum_{i=1}^{n} \beta_{ii}(t) X_i^2$
- Parameterization:
  - $\beta_q(t) = \gamma_{q,1} P_0(t) + \gamma_{q,2} P_1(t) + \cdots \gamma_{q,R} P_{R-1}(t)$
    - $q = i, ij, or\ ii$ with $i, j = 1, 2, \ldots, n; j < i$
  - R(parameters) < K(Data per Experiment)
- **DRSM-1: Parametrization with t** ➡ Has Oscillations
- **DRSM-2: Parametrization with** $\theta = \left\{1 - \exp(-\frac{t}{t_0})\right\}$
  - $0 \leq t < \infty \iff 0 \leq \theta < 1$

NO Oscillations - Excellent Model

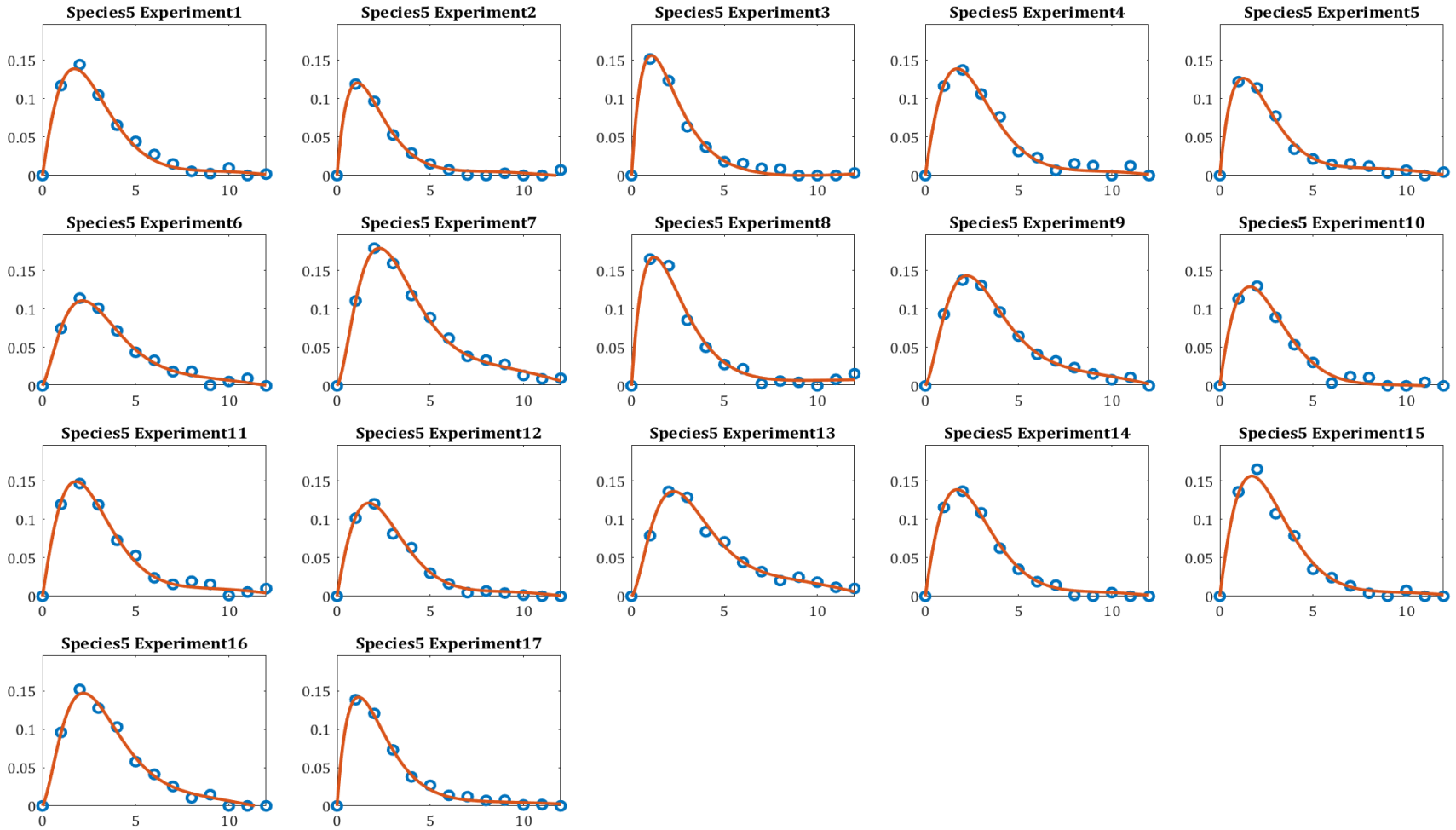
DRSM-2c for **ALL** Pfizer Data: $C_1(t)$

- Time-Resolved **Species 1** Measurements with
- $y(t) = \beta_0(t) + \sum_{i=1}^{n} \beta_i(t) X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{ij}(t) X_i X_j + \sum_{i=1}^{n} \beta_{ii}(t) X_i^2$
  - $R = 3, t_c = 3.3$ All 17 experiments



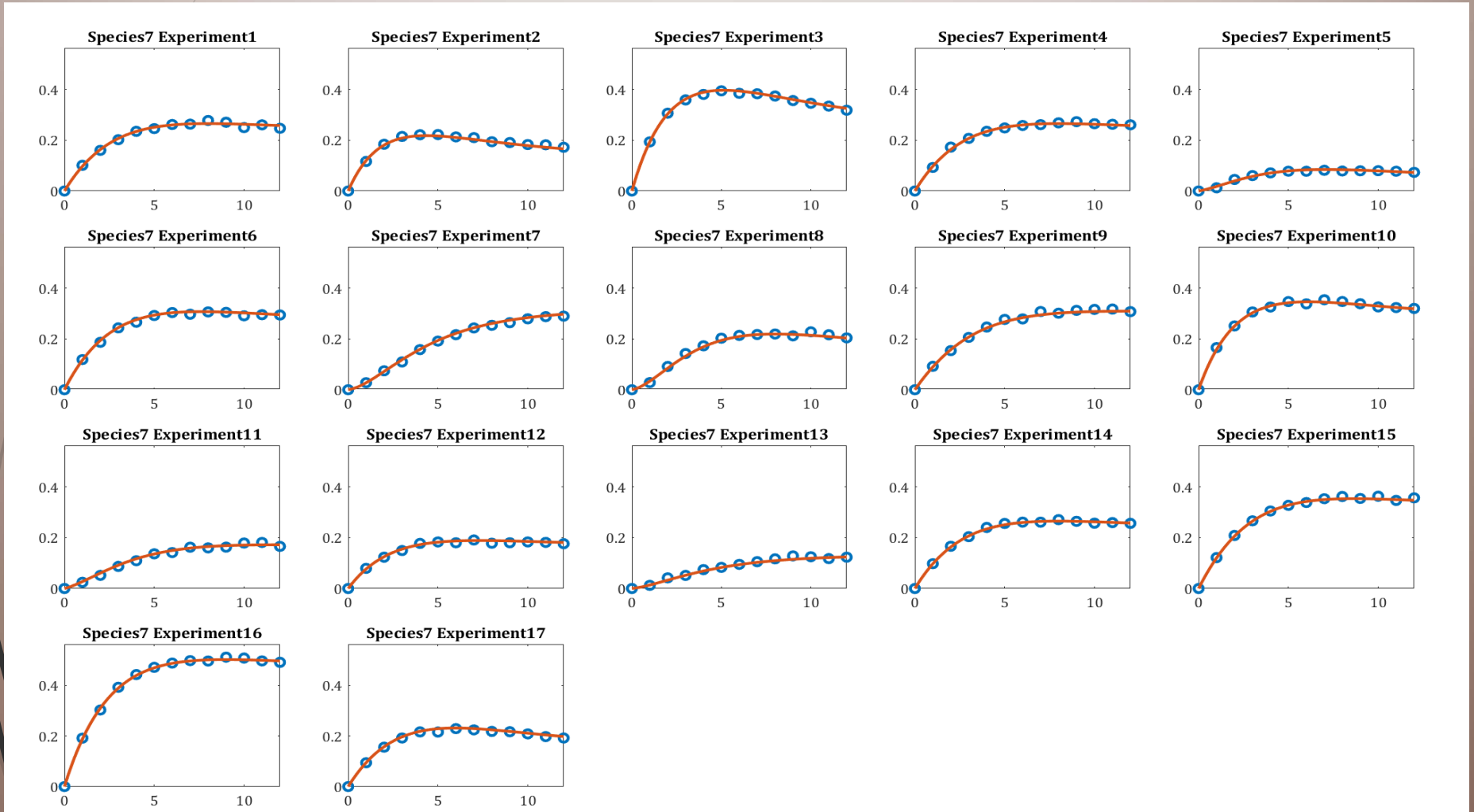
DRSM-2c for **ALL** Pfizer Data: $C_5(t)$

- **Species 5:** $R = 5, t_c = 5.4$



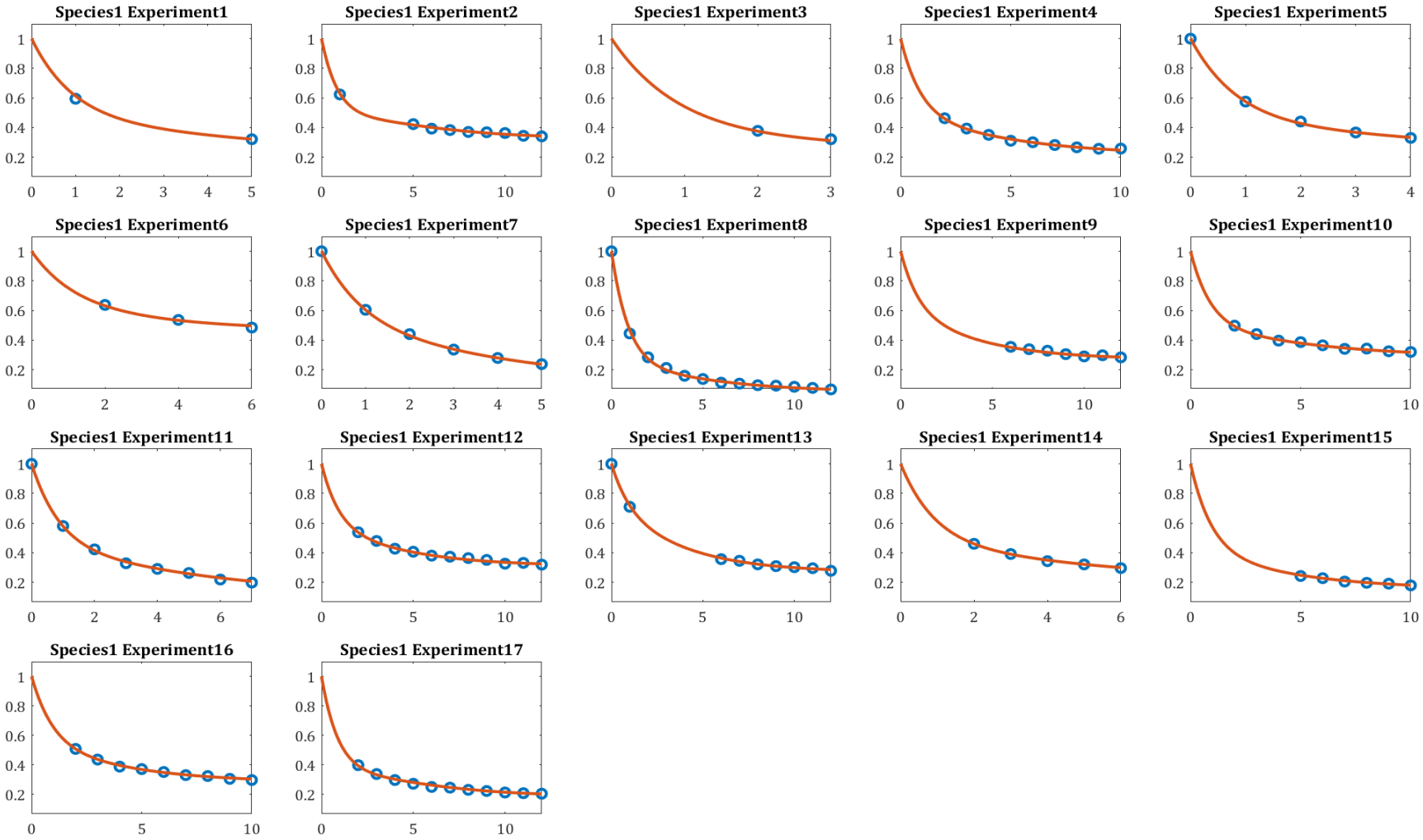
DRSM-2c for **ALL** Pfizer Data: $C_7(t)$

- **Species 7:** $R = 3, t_c = 5.4$



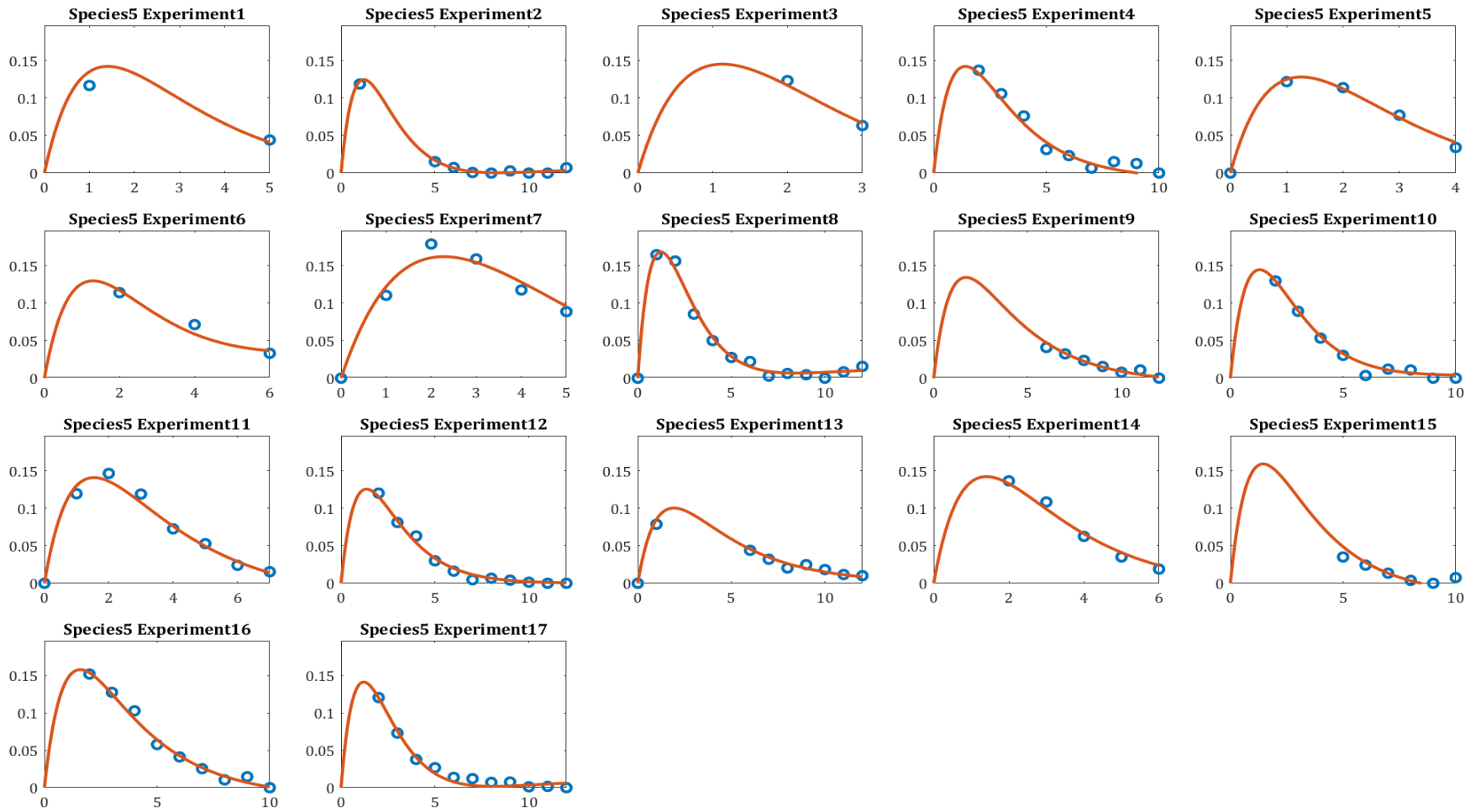
DRSM-2c: **Missing** Pfizer Data: $C_1(t)$

- Species 1: $R = 3, t_c = 3.3$



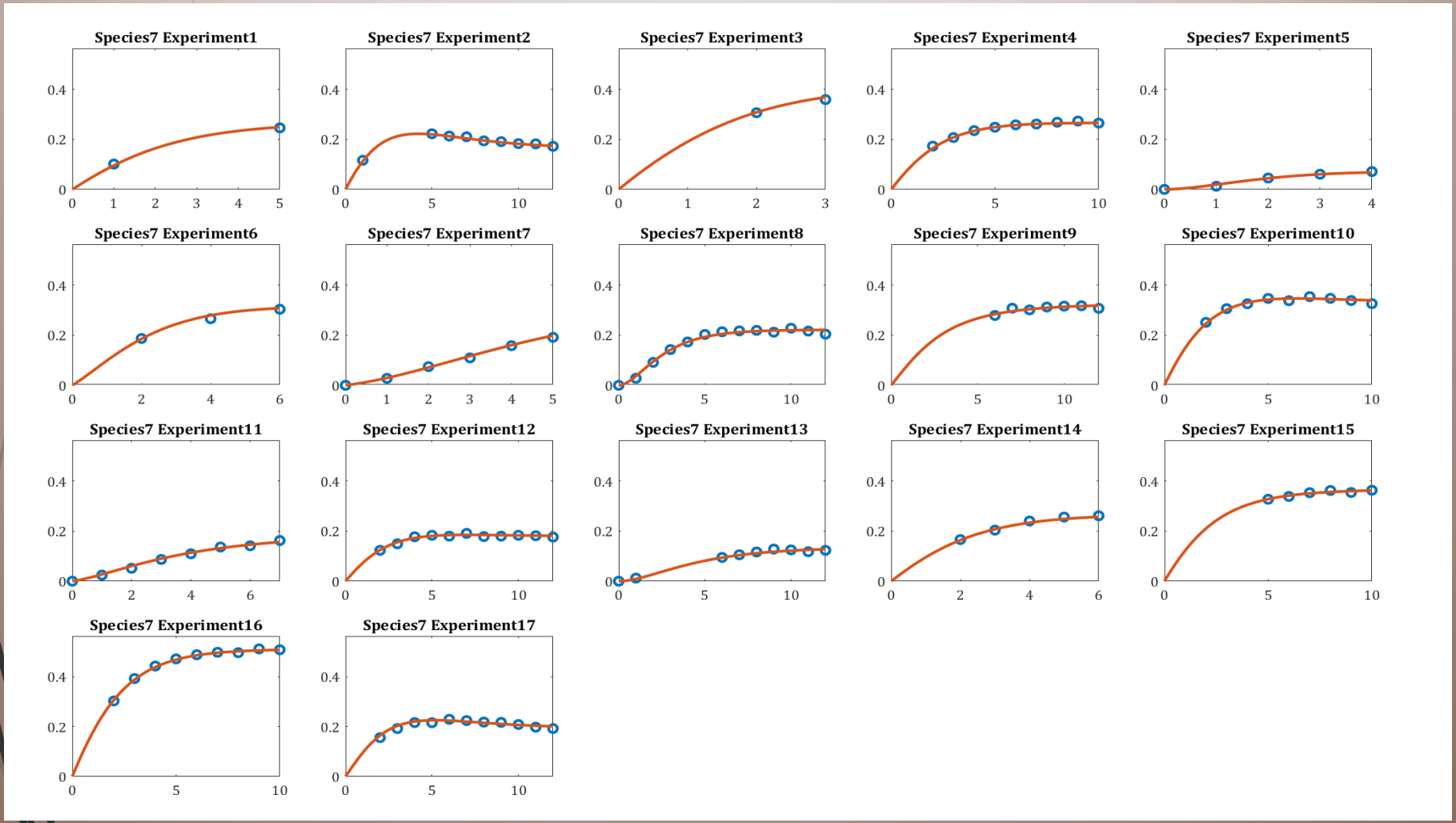
DRSM-2c: **Missing** Pfizer Data: $C_5(t)$

- Species 5: R=3 Tc=3.3

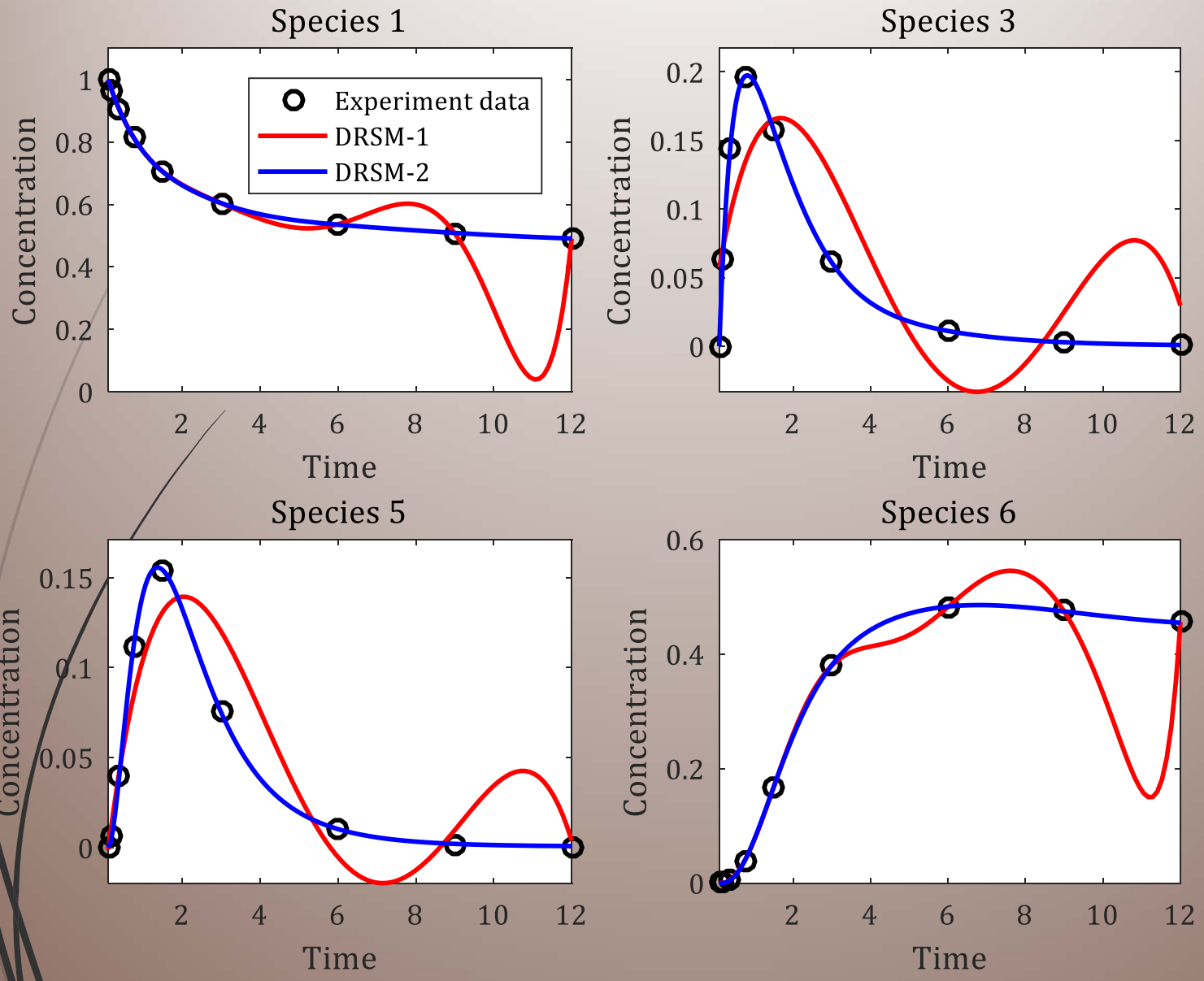


DRSM-2c: **Missing** Pfizer Data: $C_7(t)$

- Species 7: R=3 Tc=3.3

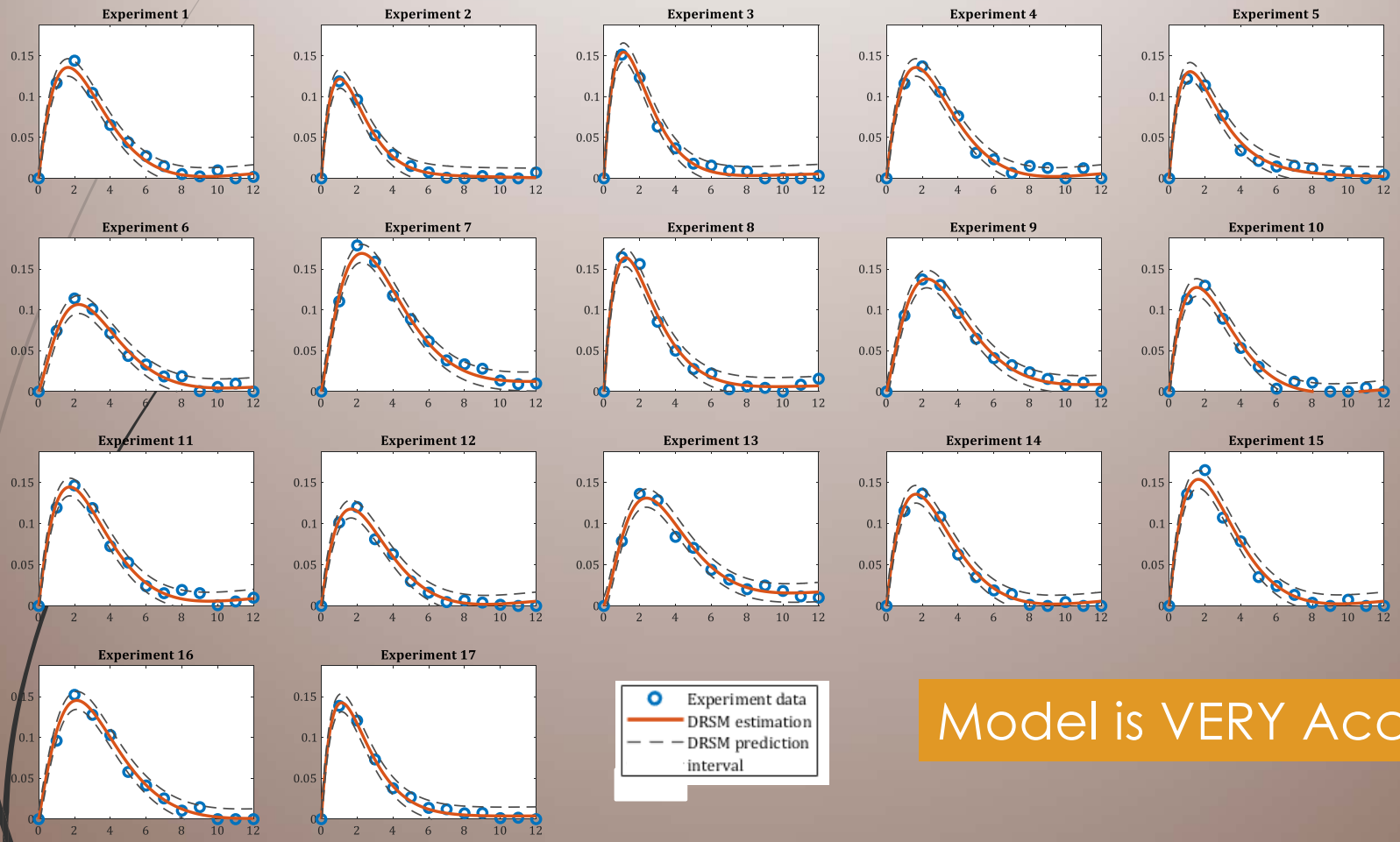


DRSM-2 vs DRSM-1 - Non-equidistant Data



VERY Notable Improvement

Species E with Prediction Interval



Model is VERY Accurate

Fractional Factorial Design (Merck)

- 3 Species and 5 Factors: A, B, C, D, and E
- 2 Blocks: Robotic & Manual
- 6 Data/Batch at ***Unequal*** Intervals

  0, 20, 40, 60, 120, 240 mins
- LC area converted to concentration

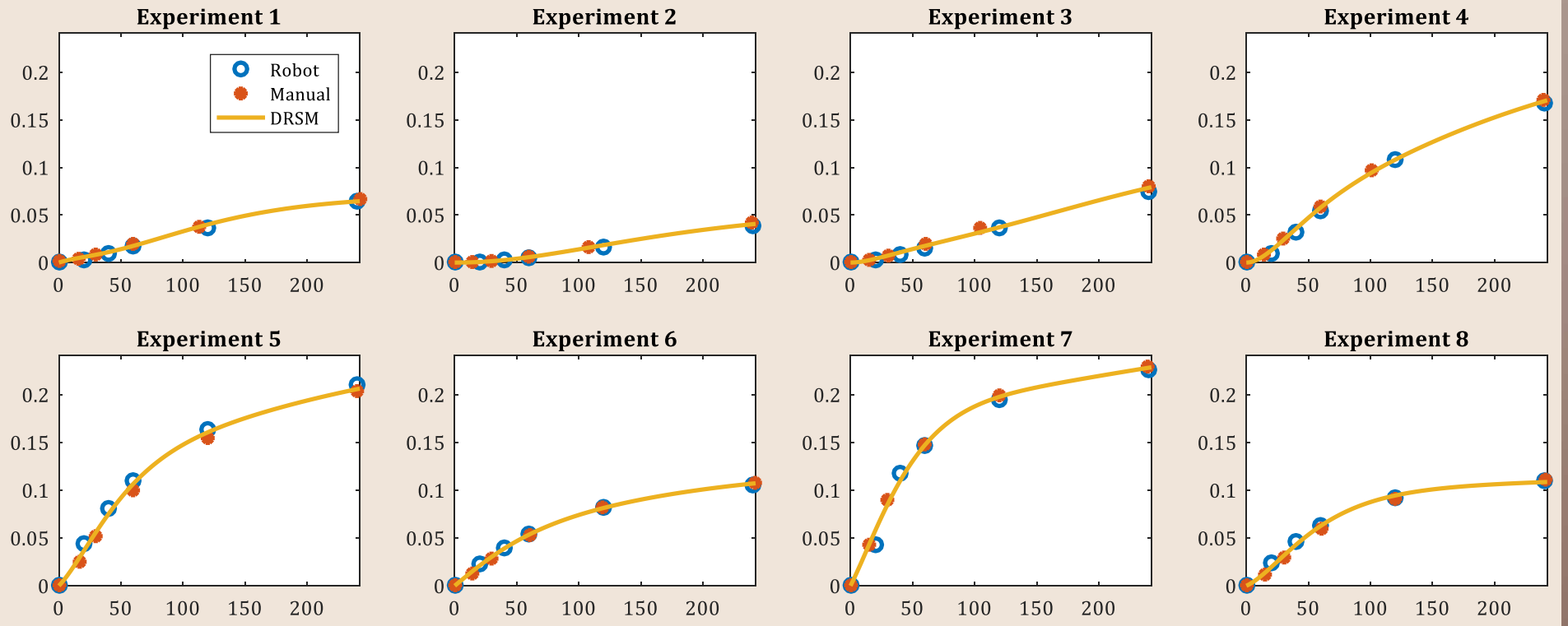
The 5 FACTORS

A: Methanol ratio, (% wt/wt solvent)
B: Starting material, wt% **C**: Base, wt%
D: Water wt% **E**: Temperature

- ¼ **Fractional factorial design**: 2^{5-2} design
  - 8 experiments
  - Aliasing Structure: D = AB, and E = AC

2FI Model: Species B

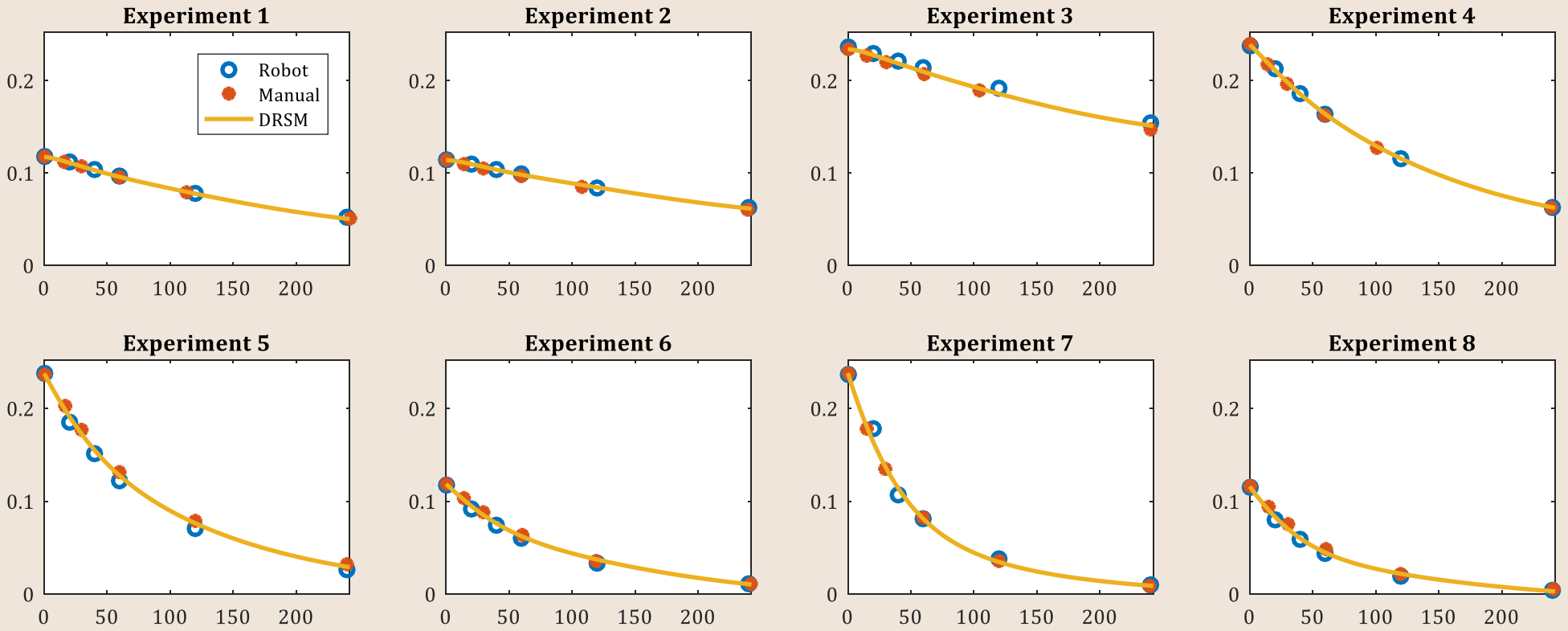
LoF *p*-value = 0.99 ➔ Perfect model



$$\tilde{y}(t) = \beta_0(t) + \beta_A(t)A + \beta_B(t)B + \beta_C(t)C + \beta_D(t)D + \beta_E(t)E +$$
$$+\boldsymbol{\beta_{BC}(t)BC + \beta_{CD}(t)CD}$$

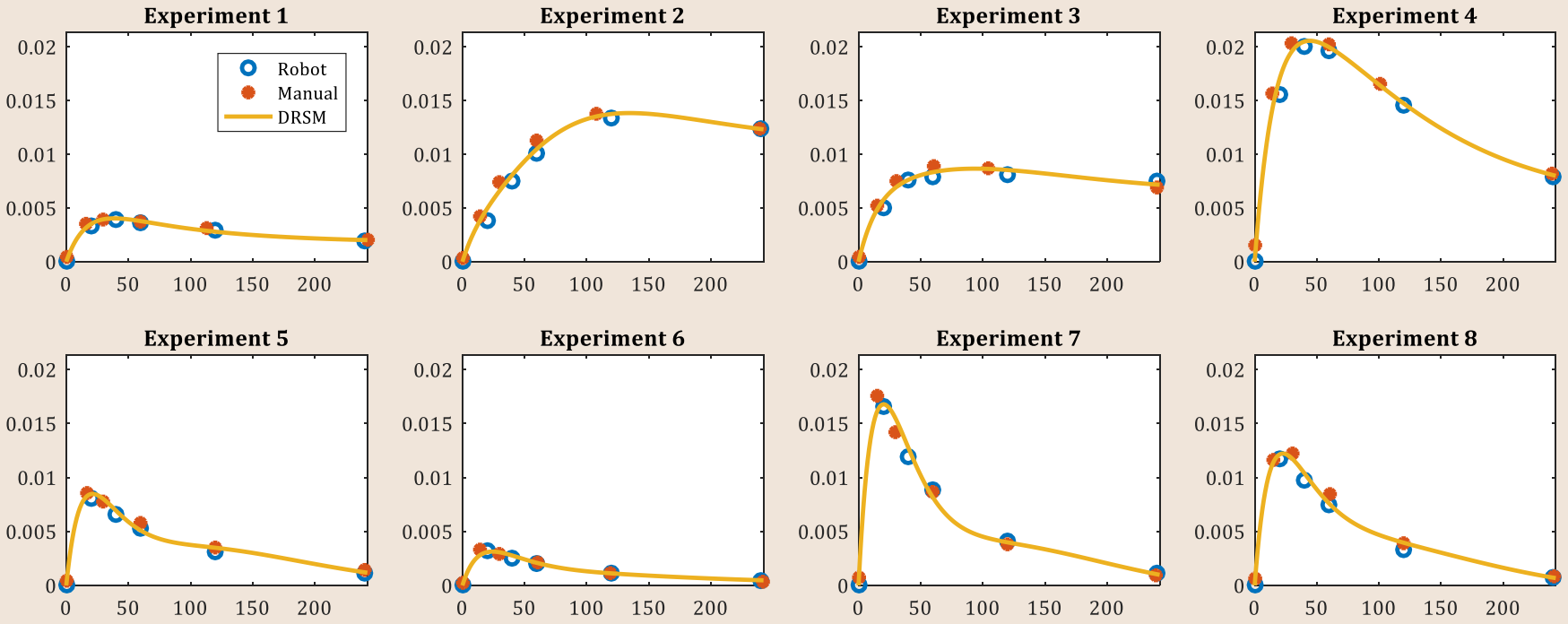
2FI Model: Species A

LoF *p*-value = 0.99 ➔ Perfect model



2FI Model: Species C

LoF *p*-value = 0.06 ➔ Good Model



$$\tilde{y}(t) = \beta_0(t) + \beta_A(t)A + \beta_B(t)B + \beta_C(t)C + \beta_D(t)D + \beta_E(t)E +$$
$$+\boldsymbol{\beta_{BC}(t)BC + \beta_{CD}(t)CD}$$

Block Effect Insignificant: Robotic vs. Manual Operation

DRSM ⟹ Reaction Knowledge

from **Data** to … **Knowledge**

From DRSM Models DISCOVER Stoichiometry and Kinetics

- Simple Semi-Batch Reactor Example → $\begin{cases} R1:\ A + B \rightarrow C \\ R2:\ 2B \rightarrow D \\ R3:\ C \rightarrow E \end{cases}$
  - Five DRSMs for $C_A(t), \ldots, C_E(t)$,

$$y_A(t) = \beta_{A0}(t) + \sum_{i=1}^{n} \beta_{Ai}(t) X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{Aij}(t) X_i X_j + \sum_{i=1}^{n} \beta_{Aii}(t) X_i^2$$

$$\vdots$$

$$y_E(t) = \beta_{E0}(t) + \sum_{i=1}^{n} \beta_{Ei}(t) X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{Eij}(t) X_i X_j + \sum_{i=1}^{n} \beta_{Eii}(t) X_i^2$$

- Calculate Derivatives with Time for ALL Models

$$y'_A(t) = \beta'_{A0}(t) + \sum_{i=1}^{n} \beta'_{Ai}(t) X_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta'_{Aij}(t) X_i X_j + \sum_{i=1}^{n} \beta'_{Aii}(t) X_i^2$$

Rate Data ⟹ Stoichiometry

- Rates of appearance for Each Species

$$\rightarrow \begin{cases} Rxn1:\ A + B \rightarrow C \\ Rxn2:\ 2B \rightarrow D \\ Rxn3:\ C \rightarrow E \end{cases}$$

- $\boldsymbol{D}_k = \begin{pmatrix} r_{Ak}(t_1) & r_{Bk}(t_1) & r_{Ck}(t_1) & r_{Dk}(t_1) & r_{Ek}(t_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{Ak}(t_i) & r_{Bk}(t_i) & r_{Ck}(t_i) & r_{Dk}(t_i) & r_{Ek}(t_i) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{Ak}(t_{n_K}) & r_{Bk}(t_{n_K}) & r_{Ck}(t_{n_K}) & r_{Dk}(t_{n_K}) & r_{Ek}(t_{n_K}) \end{pmatrix}$

$$t_i = i\Delta t \quad i = 1, \dots, n_K \quad \Delta t = 1/n_K$$

  - For $n_K = 100$ matrix D_k is a 100×5
- Data Matrix for Rates of ALL Species and ALL Experiments: R_c
  - $\boldsymbol{R}_c = \begin{pmatrix} \boldsymbol{D}_1 \\ \boldsymbol{D}_2 \\ \vdots \\ \boldsymbol{D}_{n_e} \end{pmatrix}$ D_k = Data from k−th experiment $k = 1, 2, \dots, n_e$
  - For $n_e = 9$ experiments R_c is a 900×5 matrix
- SVD=Singular Value Decomposition of $\boldsymbol{R}_c$

Number of Significant SVs = ?

  - $\boldsymbol{R}_c = \boldsymbol{U\Sigma V}^{\mathrm{T}},\ \Sigma = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_5 \end{pmatrix}$, $\boldsymbol{U\Sigma V}^{\mathrm{T}}$ $sizes: 900X5, 5X5$ & $5X5$

SVD: $\boldsymbol{R_c = U\Sigma V^{\mathrm{T}}}$ & Projections

- # of Reactions ⇔ Significant σ_i Values= 3
  - $\boldsymbol{R_c = U_3 \Sigma_3 V_3^T}$ $\quad \boldsymbol{V_3^T} = \begin{pmatrix} \boldsymbol{v_1^T} \\ \boldsymbol{v_2^T} \\ \boldsymbol{v_3^T} \end{pmatrix} = \begin{pmatrix} 0.41 & 0.84 & -0.26 & -0.21 & -0.15 \\ -0.26 & 0.21 & 0.79 & -0.23 & 0.50 \\ 0.60 & -0.28 & 0.01 & 0.44 & -0.61 \end{pmatrix}$
- **IS (-1,-1,1,0,0) a Linear Combination of the $\boldsymbol{V_3^T}$ rows ?**
- Projection Matrix: $P = \boldsymbol{V_3^T V_3}$
- Projection of Candidate Stoichiometry: $\boldsymbol{n_{ir} = n_i V_3^T V_3}$
- Is it TRUE that: $\boldsymbol{n_{ir} \cong n_i}$?
- Projection Score: $PS = 100\{1 - \|\boldsymbol{n_{ir} - n_i}\|/\|\boldsymbol{n_i}\|\}$
  - **PS ≥ 90 is GOOD**

Initial & Sequential Projections

- **PREPARATION** of Data
  - Rate Data Matrix $\boldsymbol{R}_c$ of size $n_d \times n_s$
    - n_d = # of Data, n_s = # of species
  - Number of Significant Singular Values (sSVs)
    - Statistical Determination via an *F*-test = n_{ssvs}
      - Malinowski, *J. of Chemometrics.* 1989; **3**(1):49-60
  - Define Candidate Stoichiometries
- **INITIAL** Projection Step
  - Calculate Projection Scores (PC)
    - Accept n_i Reactions with PC≥90
  - Subtract from Rate Data Contribution of Identified Rxns
- **SEQUENTIAL** Projection Step: Repeat above

Identifying Pfizer Stoichiometries

Additive error = 0.005 on Concentrations $0.005 < C_i(t_k) < 0.9$

Scores of **True** reactions		
1	A + B⇆ C + D	96.5
2	C → D + E	90.8
3	E → F	92.3
4	B + D ⇆ G	99.1
5	G → D + H	96.3
6	A + F → I	**82.4**
7	2A → J	**77.4**
8	B + J → 2E + I	**24.8**

Scores of **Untrue** reactions		
1	A → J	**57.6**
2	C → J	**38.8**
3	2A + B → J	**72.5**
4	J → 2D + I	**65.0**
5	B + J → E + I	**21.2**
6	B + J → D + I	**51.2**

Blind Test: Excellent Result

$\text{Score}_i = 100(1 - \|\boldsymbol{n}_{ir} - n_i\|/\|\boldsymbol{n}_i\|)$

n_i= Candidate Stoichiometry

$\boldsymbol{n}_{ir}$= Response Vector

Seven (7) Significant SVs via an *F*-test

σ_i = **81, 9.7, 6.3, 1.5, 1.0, 0.92, 0.22**, 0.18, 0.15, 0.09

Identifying Pfizer Stoichiometries

NO Measurement error

Scores of True reactions (Without Measurements Error)		
1	A + B ⇆ C + D	99.5
2	C → D + E	99.0
3	E → F	99.5
4	B + D ⇆ G	99.9
5	G → D + H	99.5
6	A + F → I	99.9
7	2A → J	97.9
8	B + J → 2E + I	92.7

Scores of Untrue reactions (Without Measurement error)		
1	A → J	74.6
2	C → J	57.3
3	2A + B → J	81.9
4	J → 2D + I	82.0
5	B + J → E + I	85.8
6	B + J → D + I	88.6

Confirmation of Method

Eight (8) Significant SVs

$\sigma_i =$ **81, 9.7, 6.3, 1.5, 1.0, 0.91, 0.19, 0.10**, 0.04, 0.02

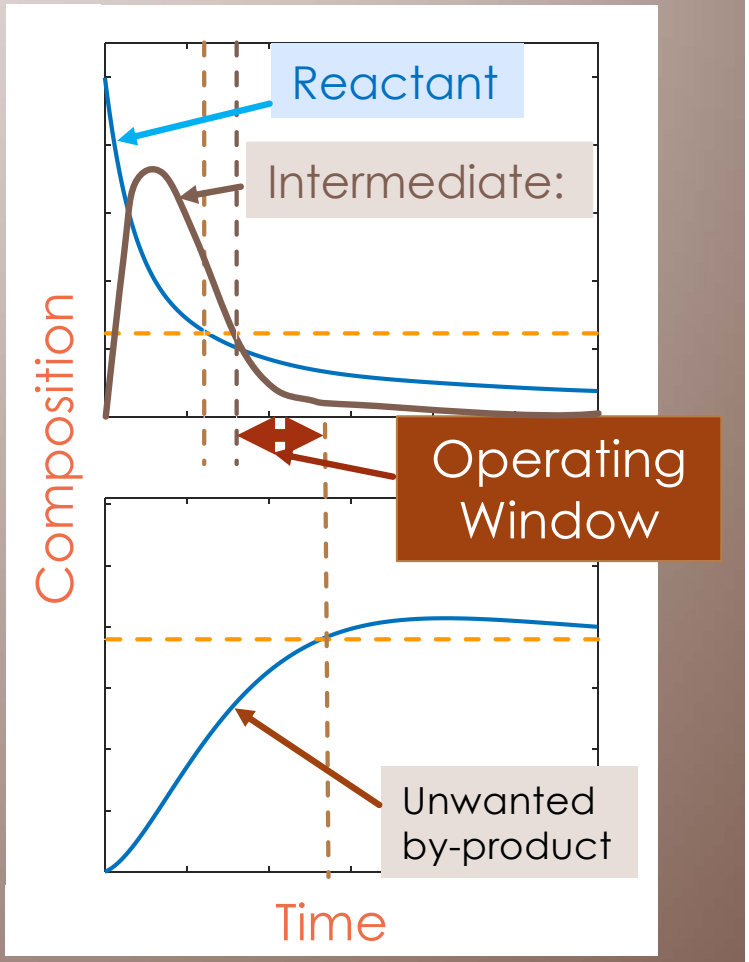
Dw

DRSM and TFA

- DRSM Makes TFA More Accurate
  - Larger Number of Significant Singular Values
    - Identifies More Reactions
- With Stoichiometries Identified
  - Calculate Rates of Each Reaction
  - Reaction Rates & $C_i(t)$ → Kinetics
    - One Reaction at a Time → MORE Accurate Model
- DRSM Enables Other Tasks
  - e.g. Optimal Operating Conditions

Process Optimization via DRSM

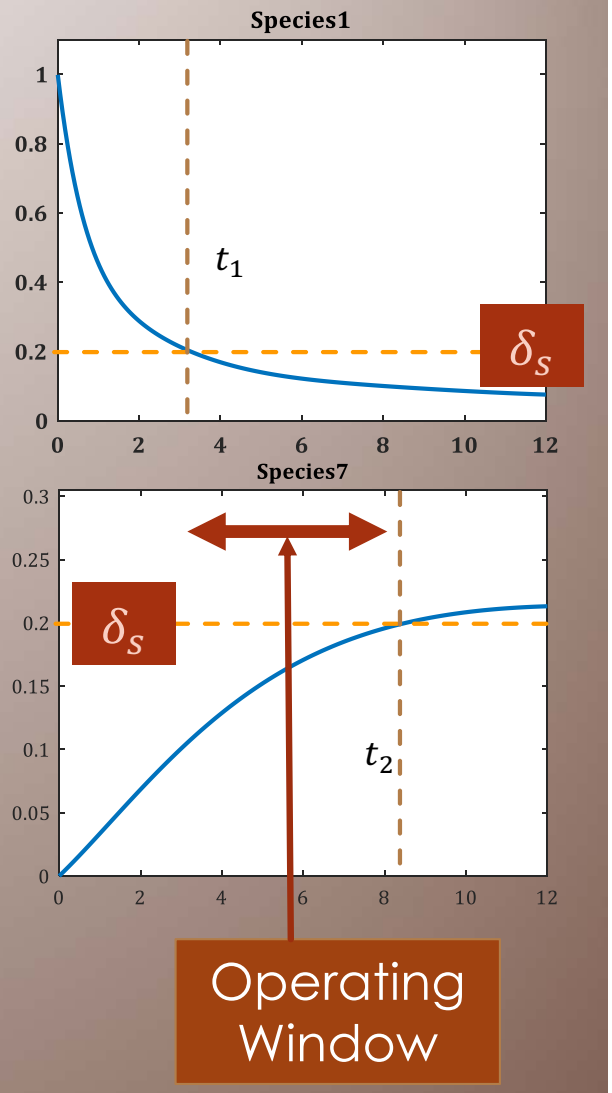
- Calculate Operating Window (OW):
  - Concentrations of Impurities Below Specs
    - Reactants
    - Intermediates
    - Unwanted by-products
- Maximize OW by Selecting
  - Operating Conditions
- Account for Uncertainties
- Peak or Area HPLC Data



Maximize Operating Window

- Optimization Results
  - When $\delta_s = 0.1$ for all species,
    - window does not exist.
  - Study different specifications

δ_s	Factor1	Factor2	Factor3	Optimal window (hr)
0.14	Infeasible			---
0.15	90	1.02	0	3.07
0.16	90	1.03	0.06	4.35
0.17	90	1.03	0.17	4.59
0.18	90	1.03	0.28	4.81
0.19	90	1.03	0.39	5.01
0.2	90	1.03	0.50	5.20



Acknowledgements

- Dr. Yachao Dong, Postdoctoral Fellow
- Mr. Jacob Santos-Marques, Tufts Graduate
- Pfizer and Merck for Financial Support
  - AND, most Importantly, an Intellectually Stimulating Collaboration
  - From Pfizer: Drs. Mustakis, Hawkins, Han, & Wang
  - From Merck: Drs. Grosser, McMullen & Stone

What to Remember Tomorrow

- The Novelty of DoDE and DRSM
  - TWO Generalizations of DoE and RSM
    - DoDE: Experiments with Time Varying Inputs
    - DRSM: Modeling Time-Varying Outputs
- Stoichiometric Identification Enhanced
  - Some Current Issues
    - Implications of Unmeasured Species
    - Not Enough Candidate Stoichiometries

Thank You Very Much

May I answer your Questions